Les 3. Steekproefonderzoek

Onderzoekstechnieken

Jens Buysse **Wim De Bruyn** **Wim Goedertier** **Bert Van Vreckem**

AJ 2018-2019

HO GENT

What's on the menu today?

Steekproefonderzoek

Kansverdeling van een steekproef

De Centrale Limietstelling

Van steekproef naar populatie
- Betrouwbaarheidsintervallen
- Betrouwbaarheidsinterval grote steekproef
- Kleine steekproef
- Betrouwbaarheidsinterval voor fractie

HO GENT

Steekproefonderzoek

HO GENT

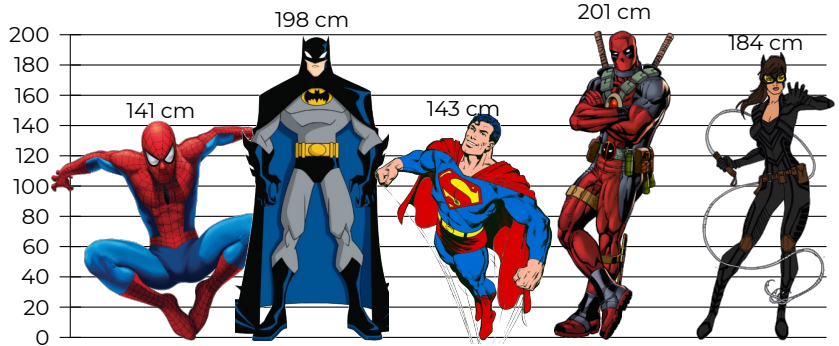
USA Today has come out with a new survey. Apparently, three out of every four people make up 75% of the population

—David Letterman

HO
GENT

Steekproef en Populatie

Herinner je onze superhelden



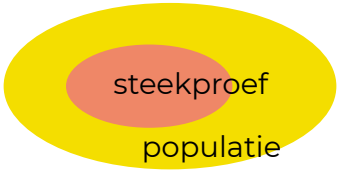
HO GENT

Steekproef en Populatie

Steekproef en Populatie

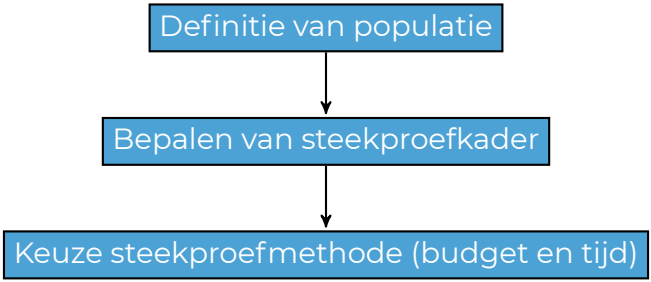
De verzameling van alle objecten of personen waar men in geïnteresseerd is voor het onderzoek noemt men de populatie.

Wanneer men een subgroep uit een populatie gaat onderzoeken, dan noemen we die groep een steekproef.



HO GENT

Methode om tot een steekproef te komen



HO GENT

Gestratificeerd naar variabelen

	Leeftijd				
Geslacht	≤ 18	$]18, 25]$	$]25, 40]$	> 40	Totaal
Vrouw	500	1500	1000	250	3250
Man	400	1200	800	160	2560
Totaal	900	2700	1800	410	5810

HO GENT

Gestratificeerd naar variabelen

	Leeftijd				
Geslacht	≤ 18	$]18, 25]$	$]25, 40]$	> 40	Totaal
Vrouw	500	1500	1000	250	3250
Man	400	1200	800	160	2560
Totaal	900	2700	1800	410	5810

	Leeftijd				
Geslacht	≤ 18	$]18, 25]$	$]25, 40]$	> 40	Totaal
Vrouw	50	150	100	25	325
Man	40	120	80	16	256
Totaal	90	270	180	41	581

HO GENT

Hoe elementen voor een steekproef kiezen?

Aselecte (of random) steekproef : elk element uit de onderzoekspopulatie heeft een even grote kans om in de steekproef terecht te komen

Selecte steekproef : de elementen voor de steekproef worden *niet* random geselecteerd. Objecten die *gemakkelijk* kunnen verzameld worden, hebben meer kans om in de steekproef terecht te komen. (convenience sampling)

HO
GENT

Mogelijke fouten

Metingen in een steekproef zullen typisch afwijken van de waarde in de hele populatie $\Rightarrow$ Fouten!

- Toevallig $\leftrightarrow$ Systematisch
- Steekproeffout $\leftrightarrow$ Niet-steekproeffout

HO GENT

Steekproeffouten

- Toevallige steekproeffouten
  - Puur toeval

HO GENT

Steekproeffouten

- Toevallige steekproeffouten
  - Puur toeval
- Systematische steekproeffouten
  - Online enquête: mensen zonder internet worden uitgesloten
  - Straatenquête: enkel wie daar op dat moment loopt
  - Vrijwillige enquête: enkel geïnteresseerden doen mee

HO GENT

Niet-steekproeffouten

- Toevallige niet-steekproeffouten
  - Verkeerd aangekruiste antwoorden

HO GENT

Niet-steekproeffouten

- Toevallige niet-steekproeffouten
  - Verkeerd aangekruiste antwoorden
- Systematische niet-steekproeffouten
  - Slechte of niet geijkte meetapparatuur (slechte weegschaal)
  - Waarde kan beïnvloed worden door het feit dat je meet
  - Respondenten liegen (aantal sigaretten per dag)

HO GENT

Variantie/standaarddeviatie van een steekproef

Aangepaste formule:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{x} - x_i)^2$$

De reden voor de wijziging kan je wiskundig bewijzen, maar we gaan het empirisch onderzoeken

R-script: cursus/data/sample-variance.R

HO GENT

Overzicht symbolen

	populatie	steekproef
aantal objecten	N	n
gemiddelde	μ	$\overline{x}$
variantie	$\sigma^2 = \frac{\sum(\mu - x_i)^2}{N}$	$s^2 = \frac{\sum(\overline{x} - x_i)^2}{n-1}$
standaarddeviatie	σ	s

HO GENT

Kansverdeling van een steekproef

HO GENT

Wat weten we nog van de kansrekening?

- Uitkomstenruimte
- Uitkomst
- Gebeurtenis
- Kansruimte

HO GENT

Kansverdeling voor één dobbelsteen

Wat is de kans om een aantal ogen te werpen met een dobbelsteen?

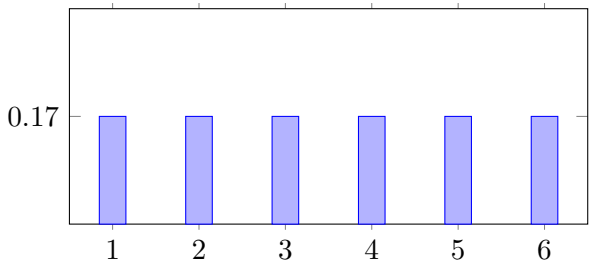
1	2	3	4	5	6

HO GENT

Kansverdeling voor één dobbelsteen

Wat is de kans om een aantal ogen te werpen met een dobbelsteen?

1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



HO GENT

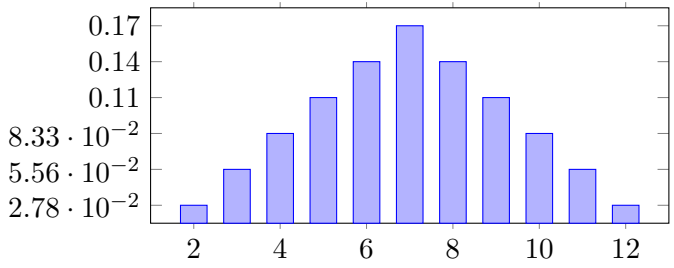
Kansverdeling voor twee dobbelstenen

2	3	4	5	6	7	8	9	10	11	12

HO GENT

Kansverdeling voor twee dobbelstenen

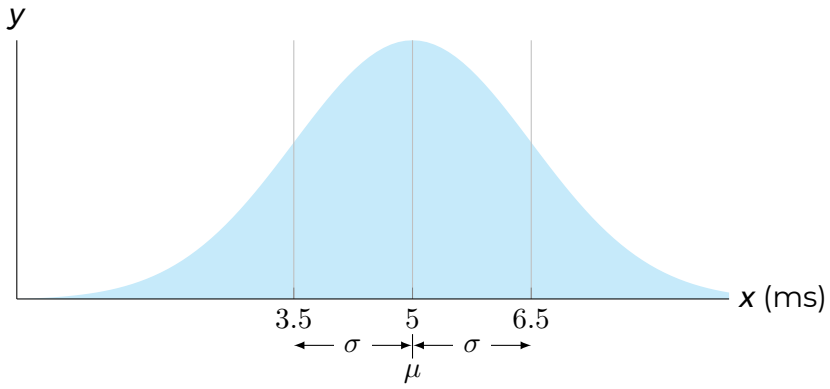
2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



HO GENT

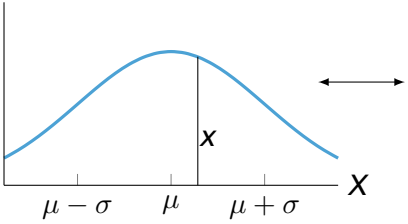
Continue kansverdeling

De reactiesnelheid *x* van Superman (in milliseconden) kun je als volgt weergegeven:



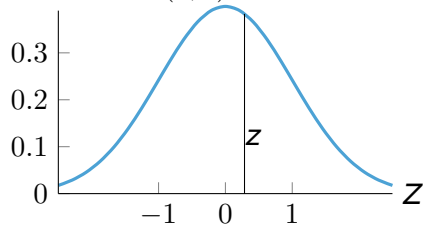
*Standaard*normale verdeling

normaalverdeling $x \in X \sim Nor(\mu, \sigma)$



standaardnormaalverdeling

$z \in Z \sim Nor(0, 1)$

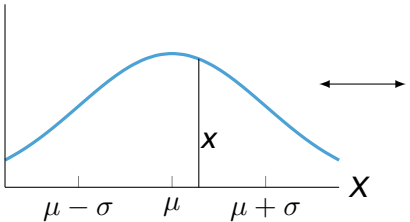


x en *z* hebben een vergelijkbare positie op de Gauss-curve.
Wat is het wiskundig verband tussen *x* en *z*?

HO GENT

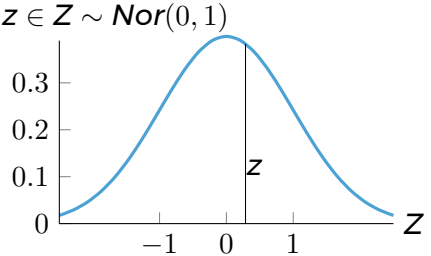
*Standaard*normale verdeling

normaalverdeling $x \in X \sim Nor(\mu, \sigma)$



standaardnormaalverdeling

$z \in Z \sim Nor(0, 1)$

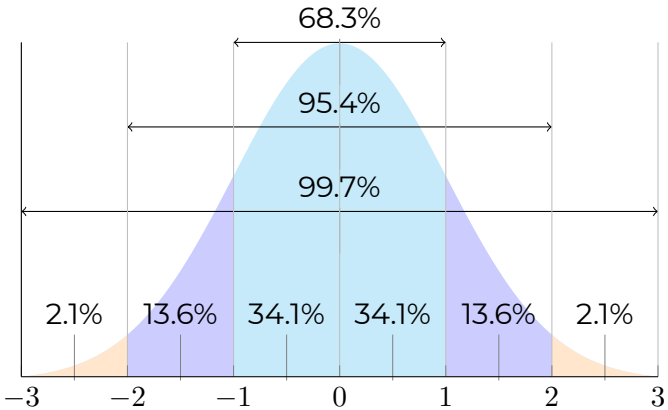


x en *z* hebben een vergelijkbare positie op de Gauss-curve.
Wat is het wiskundig verband tussen *x* en *z*?

$X = \mu + Z.\sigma$ and $Z = \frac{X-\mu}{\sigma}$

HO GENT

Standaardnormale verdeling



HO GENT

Belangrijkste functies in R

Voor een normale verdeling met gemiddelde `m` en standaardafwijking `s`:

Functie	Betekenis
`pnorm(x, m, s)`	Linkerstaartkans, $P(X < x)$
`dnorm(x, m, s)`	Hoogte van de Gausscurve op punt `x`
`qnorm(p, m, s)`	Onder welke grens zal `p`% van de waarnemingen liggen?
`rnorm(n, m, s)`	Genereer `n` normaal verdeelde random getallen

Argumenten `m` en `s` weglaten geeft waarden voor de ***standaard***normaalverdeling: `pnorm(x)=pnorm(x,0,1)`

HO GENT

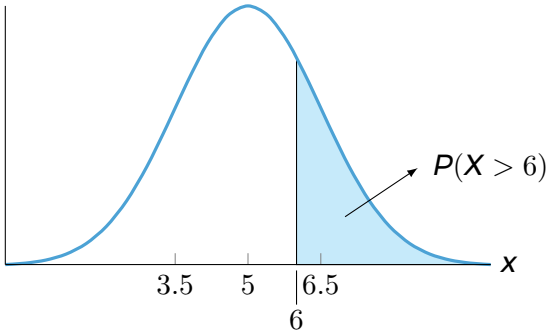
Kansen berekenen

Hoe groot is de kans dat...

...de reactiesnelheid van Superman meer dan 6 milliseconden is?

Wiskundige notatie:

$P(X > 6)$ = ? met $X \sim Nor(\mu = 5, \sigma = 1,5)$



HO GENT

Kansen berekenen: z-tabel

$P(X > 6)$ = ? met $X \sim Nor(\mu = 5, \sigma = 1,5)$

(Oude) berekeningsmethode met z-tabel, vb.
http://sixsigmastudyguide.com/wp-content/uploads/2014/04/z-table.jpg

HO GENT

Kansen berekenen: z-tabel

$P(X > 6)$ = ? met $X \sim Nor(\mu = 5, \sigma = 1,5)$

(Oude) berekeningsmethode met z-tabel, vb.
`http://sixsigmastudyguide.com/wp-content/uploads/2014/04/z-table.jpg`

1. zet om naar z-score
   $z = \frac{6-5}{1,5} = 0,667$ dus $P(X > 6) = P(Z > 0,667)$
2. zet eerst om naar een ***linker***staartkans
   - regel van 100% kans: $P(Z > 0,667) = 1 - P(Z < 0,667)$
   - of symmetrie-regel: $P(Z > 0,667) = P(Z < -0,667)$
3. zoek op in z-tabel

HO GENT

Kansen berekenen: met R

$P(X > 6)$ = ? met $X \sim Nor(\mu = 5, \sigma = 1,5)$

Rechtstreekse berekening met R (of rekenmachine)

HO GENT

Kansen berekenen: met R

$P(X > 6)$ = ? met $X \sim Nor(\mu = 5, \sigma = 1,5)$

Rechtstreekse berekening met R (of rekenmachine)

- zet eerst om naar een ***linker***staartkans:
  met regel van 100% kans: $P(X > 6) = 1 - P(X < 6)$
- bereken met R: $1 - P(X < 6) =$`1-pnorm(6,5,1.5)`

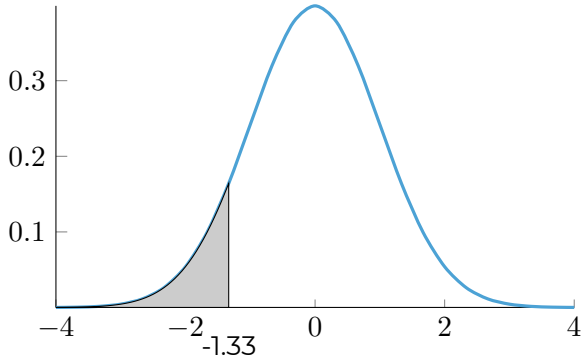
HO GENT

Voorbeelden

1. Hoe groot is de kans dat de reactiesnelheid van Superman minder dan 4 ms is?
2. Hoe groot is de kans dat hij in minder dan 7 ms reageert?
3. Hoe groot is de kans dat Superman in minder dan 3 ms reageert?
4. Hoe groot is de kans dat hij reageert tussen de 2 en de 6,5 ms?
5. Binnen welke tijd ligt 80% van zijn reactiesnelheid?

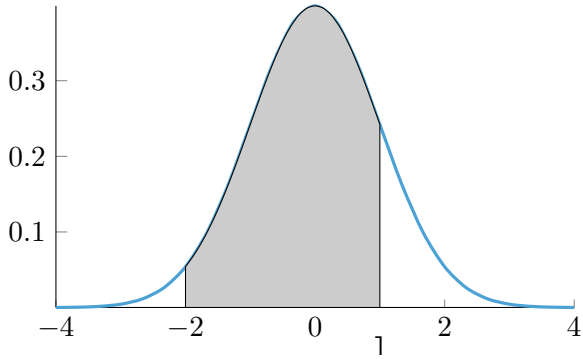
HO GENT

Vraag 3: $P(X < 3)$



HO GENT

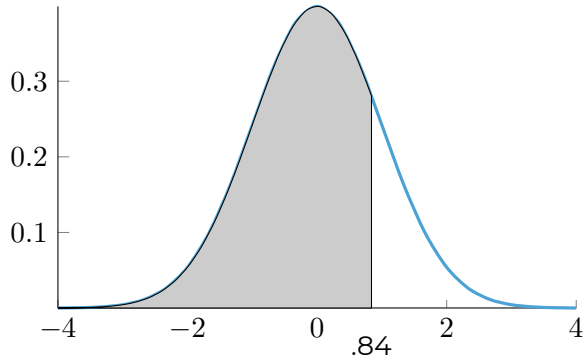
Vraag 4: $P(2 < X < 6,5)$



HO GENT

Vraag 5

Voor welke x is $P(X < x) = 80\%$?



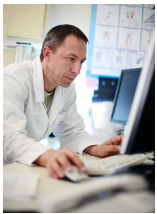
HO GENT

De Centrale Limietstelling

HO
GENT

De centrale limietstelling

Als de steekproefomvang voldoende groot is, dan kan de kansverdeling van het steekproefgemiddelde benaderd worden met een normale verdeling. Dit geldt ongeacht de vorm van de kansverdeling van de onderliggende populatie



- 1 test
- 25 tests
- 100 tests

Demo: `https://students.brown.edu/seeing-theory/probability-distributions/index.html`

HO GENT

De centrale limietstelling

Beschouw een aselecte steekproef van n waarnemingen die uit een populatie met een willekeurige distributie met verwachtingswaarde μ en standaardafwijking σ. Als n groot genoeg is zal de kansverdeling van $\overline{x}$ een normale verdeling benaderen met gemiddelde $\mu_{\overline{x}} = \mu$ en standaardafwijking $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$.

Hoe groter de steekproef is, des te beter zal de kansverdeling van $\overline{x}$ overeenkomen met een normaalverdeling.

HO GENT

Van steekproef naar populatie

HO GENT

Puntschatter

Een puntschatter voor een populatieparameter is een regel of een formule die ons zegt hoe we uit de steekproef een getal moeten berekenen om de populatieparameter te schatten.

HO GENT

Betrouwbaarheidsinterval

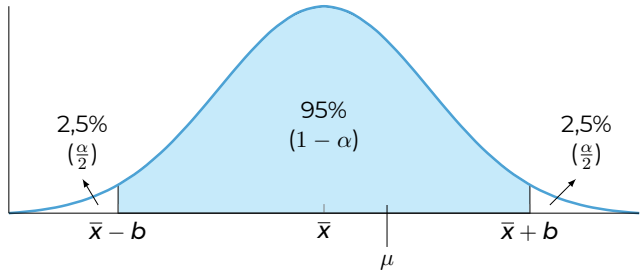
Een betrouwbaarheidsinterval is een regel of een formule die ons zegt hoe we uit de steekproef een interval kunnen berekenen dat de waarde van de parameter met een zeker betrouwbaarheidsniveau bevat.

HO GENT

Grote steekproef

Gegeven een steekproef met gemiddelde $\overline{x}$.
We zoeken nu een interval $[\,\overline{x}-b\,,\ \overline{x}+b\,]$ waarvan we met een betrouwbaarheidsniveau $(1-\alpha)$ van bvb. 95% kunnen zeggen dat μ erbinnen ligt.

$$P\,(\overline{x}-b<\mu<\overline{x}+b)=1-\alpha=0,95$$



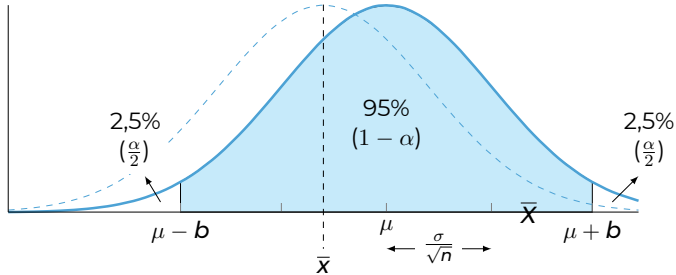
HO GENT

Grote steekproef

Dankzij de centrale limietstelling weten we dat $\overline{x} \in \overline{X} \sim Nor(\mu, \frac{\sigma}{\sqrt{n}})$

En omwille van de symmetrie weten we dat

$$P(\overline{x} - b < \mu < \overline{x} - b) = P(\mu - b < \overline{x} < \mu + b)$$



HO GENT

Grote steekproef

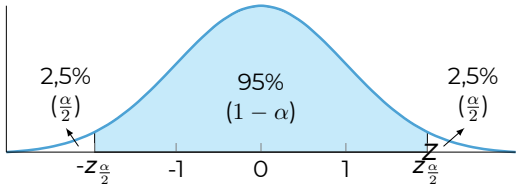
We bepalen nu de z-score voor $\overline{x}$: $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

We zoeken (in een tabel of rekentoestel) de waarde $z_{\frac{\alpha}{2}}$ waarbij:

$$P\left(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha = 0.95$$

$$P\left(z < z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} = 0.975$$

$$z_{\frac{\alpha}{2}} = \texttt{qnorm(0.975)} \approx 1.96$$



HO GENT

Grote steekproef

We krijgen nu:

$$P\left(-1,96 < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1,96\right) = 0,95$$

$$P\left(\mu - 1,96\frac{\sigma}{\sqrt{n}} < \overline{x} < \mu + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Wegens symmetrie kunnnen we μ en $\overline{x}$ verwisselen:

$$P\left(\overline{x} - 1,96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

We kunnen nu met 95% zekerheid zeggen dat:

$$\mu \in \left[\ \overline{x} - 1,96.\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + 1,96.\frac{\sigma}{\sqrt{n}}\ \right]$$

(in de praktijk gebruiken we $s_{steekpr}$ als schatting voor $\sigma_{populatie}$)

HO GENT

Kleine steekproef

Bij een ***kleine*** steekproef $(n < 30)$ is de centrale limietstelling **niet** meer geldig.

Er geldt wél:

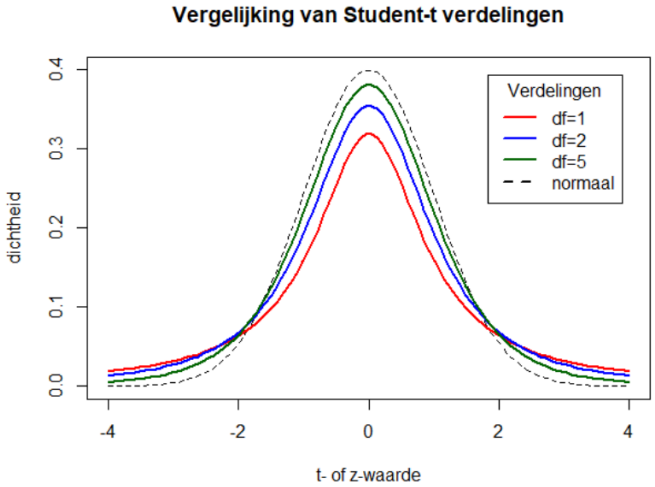
Als een populatie X normaal verdeeld is $(X \sim Nor(\mu, \sigma))$ en je neemt een ***kleine*** steekproef met gemiddelde $\overline{x}$ en standaardafwijking s, dan is

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

verdeeld volgens een t-verdeling met $n - 1$ vrijheidsgraden (degrees of freedom)

HO GENT

De Student *t*-verdeling



HO GENT

Student t-verdeling in R

Voor een t-verdeling met `df` vrijheidsgraden:
(`df = degrees of freedom`)

Functie	Betekenis
`pt(x, df)`	Linkerstaartkans, $P(X < x)$
`dt(x, df)`	Hoogte van de curve op punt `x`
`qt(p, df)`	Onder welke grens zal `p`% van de waarnemingen liggen?
`rt(n, df)`	Genereer `n` random getallen volgens deze verdeling

HO GENT

Betrouwbaarheidsinterval voor *kleine* steekproef

Om een betrouwbaarheidsinterval voor het gemiddelde μ van een populatie te bepalen op basis van een ***kleine*** steekproef, zoeken we eerst $t_{\frac{\alpha}{2},df}$.

Bij een betrouwbaarheidsniveau van 95% is $\frac{\alpha}{2} = 0,025$

Veronderstel $n = 5$ (dus `df=4`), dan is

$$t_{\frac{\alpha}{2},df} = \texttt{qt(0.975,4)} = 2.776$$

We kunnen dan met 95% zekerheid zeggen dat:

$$\mu \in \left[\overline{x} - t_{\frac{\alpha}{2},df} . \frac{s}{\sqrt{n}} \ , \ \overline{x} + t_{\frac{\alpha}{2},df} . \frac{s}{\sqrt{n}} \right]$$

HO GENT

Betrouwbaarheidsinterval voor fractie

$$\overline{p} = \frac{\text{aantal successen}}{n}$$

- Verwachting van kansverdeling van $\overline{p}$ is p.
- De standaardafwijking van kansverdeling $\overline{p}$ is $\sqrt{\frac{pq}{n}}$
- Voor grote steekproeven is $\overline{p}$ bij benadering normaal verdeeld.

Aangezien $\overline{p}$ een steekproefgemiddelde is van het aantal successen, stelt dit ons in staat een betrouwbaarheidsinterval te berekenen analoog als die voor de intervalschatting van μ voor grote steekproeven.

$$\overline{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{pq}}{n}}$$

met $\overline{p} = \frac{x}{n}$ en $\overline{q} = 1 - \overline{p}$

HO GENT